A significance test for forward stepwise model selection

Jonathan Taylor* **and Joshua Loftus and ???**

Department of Statistics
Stanford University
Sequoia Hall
390 Serra Mall
Stanford, CA 94305, U.S.A.
e-mail: jonathan.taylor@stanford.edu

Abstract: We apply the methods of Taylor et al. (2013) and Lockhart et al. (2013) on significance tests for penalized regression to forward stepwise model selection. For the kth variable to enter the model, previous work relied on an asymptotic null distribution that breaks down when grouped variables are entering the model and is difficult to derive. We iteratively apply an exact null distribution for the first (potentially grouped) variable on the residual from previous steps. The resulting method has the strengths of stepwise selection, for example parallel computation, but also remedies the problem of inflated test statistics and over-fitting.

AMS 2000 subject classifications: Primary 62M40; secondary 62H35.
Keywords and phrases: penalized regression, convex analysis, least squares, Gaussian processes.

1. Introduction

Forward stepwise regression is a stochastic model selection procedure that begins with an empty model and adds the best predictor variable in each step. Classical significance tests fail when a model has been selected this way and tend to be anti-conservative. Recently, Lockhart et al. (2013) found a novel test statistic with an appropriate null distribution that behaves well when a model has been selected using the lasso (Tibshirani, 1996). Taylor et al. (2013) modified and extended those results to the group lasso (Ming and Lin, 2005) and other adaptive regression problems. The present work explores the behavior of those test statistics for models selected by forward stepwise procedures and works out some of the details involved in applying these methods to models with grouped variables. Our test statistic can be used for valid significance testing when computed from the same data as the model selection. The resulting method can be more statistically efficient than validation on held-out data, and also more computationally efficient than penalized methods with regularization parameters chosen by cross-validation.

(**To do:** Change this paragraph as the sections are completed) In Section 2 we establish notation and describe our forward stepwise procedure. Section

*Supported in part by NSF grant DMS 1208857 and AFOSR grant 113039.

3 briefly reviews the parts of Lockhart et al. (2013) and Taylor et al. (2013) relevant to our significance test, and describes the group lasso which we require for applying the test with grouped variables. Simulation results in Section 4 using various stopping rules for the forward stepwise procedure–including some from **?**–appear promising.

2. Forward stepwise model selection

We allow forward stepwise selection to add groups of variables in each step, not only in the case of binary encoding for categorical variables but also for any grouping purpose. For example, groups of variables may be pre-designated factors such as expression measurements for all genes in a single regulatory pathway. An entire group is included or excluded together from the final model. For consistency we will use g, h as indices rather than the usual i, j throughout. Since single variables can be considered groups of size 1, our general exposition includes non-grouped situations as a special case.

Label the outcome variable of n i.i.d. measurements $y \in \mathbb{R}^n$. Let an integer $G \geq 1$ be the number of groups. For each $1 \leq g \leq G$ the design matrix encoding the gth group is the $n \times p_g$ matrix denoted X_g, where p_g is the number of individual variables in group g. Define $p = \sum_{g=1}^{G} p_g$ be the total number of individual variables, so $p = G$ in the case where all groups have size 1. Let X be the matrix constructed by column-binding the X_g, that is

$$X = \begin{pmatrix} X_1 & X_2 & \cdots & X_G \end{pmatrix}$$

With each group we associate the $p_g \times 1$ coefficient vector β_g, and write β for the $p \times 1$ vector constructed by stacking all of the β_g in order. Finally, our model for the response is

$$\begin{aligned} y &= X\beta + \sigma\epsilon \\ &= \sum_{g=1}^{G} X_g \beta_g + \sigma\epsilon \end{aligned} \tag{1}$$

where ϵ is noise. Unless otherwise specified we assume i.i.d. Gaussian noise $\epsilon \sim N(0, I_{n \times n})$ and that σ is known.

Before we describe the forward stepwise procedure we require one last ingredient. To each group of variables we assign a weight w_g. These weights act like penalties or costs, so increasing w_g makes it *more difficult* for the group X_g to enter the model. The modeler can choose weights arbitrarily, but we will only use one particular choice, based on p_g, that we discuss later. With this we are ready to describe the forward stepwise procedure Algorithm 1.

To do: Expand this section

- Brief background, cite some original papers
- Stopping rules, cite new paper on lasso?

Data: An n vector y and $n \times p$ matrix X of G grouped variables
Result: Active set A of variable groups included in the model

$A \leftarrow \emptyset$
$A^c \leftarrow \{1, \ldots, G\}$
$r_0 \leftarrow y$
for *counter* $\leftarrow 1$ **to** $\min(n, G)$ **do**
  $g^* \leftarrow \text{argmax}_{g \in A^c} \|X_g^T r_{g-1}\|_2 / w_g$
  $A \leftarrow A \cup \{g^*\}$
  $A^c \leftarrow A^c \backslash \{g^*\}$
  $r_g \leftarrow$ `lsfitResidual`(r_{g-1}, X_{g^*})
end
return A

Algorithm 1: Forward stepwise procedure

3. Significance testing: from lasso to group lasso

To do: Update this section to match/complement (Taylor et al., 2013)

In a recent work Lockhart et al. (2013) defined a *covariance test statistic* for testing the significance of a variable entering the model along the lasso solution path. They derived a simple asymptotic null distribution, proved a type of convergence under broad "minimum growth" conditions, and demonstrated in simulations that the test statistic closely matches its asymptotic distribution even in finite samples. That work marked an important advance in the problem of combining inference with model selection. The current paper extends some of their results to the group lasso (ref Yuan and Lin?). In the process we had to derive an exact finite sample null distribution for the test statistic (ref TLTT?). We also show that the techniques used to get these results can be used to do a forward stepwise procedure related to the group lasso that adds groups of variables (or factors) in each step.

3.1. Background

In the ordinary least squares setting, a significance test for a single variable can be conducted by comparing the drop in residual sums of squares (RSS) to a χ^2_1 distribution. Similarly, when adding a group of k variables we can compare the drop in RSS to a χ^2_k random variable. This generally does not work when the variable to be added has been chosen by a method that uses the data, and in particular it fails for forward stepwise procedures which add the "best" (e.g. most highly correlated) predictor in each step. In that case, the test statistic (drop in RSS) does not match the theoretical null distribution even when the null hypothesis is true. Lockhart et al. (ref LTTT) introduced a new test statistic based on the knots in the lasso solution path. Writing $\hat{\beta}(\lambda)$, the solution for a fixed value of λ, we need the following facts (ref Tibs2012)

- The vector valued function $\hat{\beta}(\lambda)$ is a continuous function of λ. For the lasso path, the coordinates of $\hat{\beta}(\lambda)$ are piecewise linear with changes in

slope occurring at a finite number of λ values referred to as *knots*. The knots depend on the data and are usually written in order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$. We follow this convention.

- The *active set* A_k is a set of indices of variables for which the corresponding coordinates of $\hat{\beta}(\lambda_k)$ are potentially nonzero. Any variable with index not in A_k has a zero coefficient in $\hat{\beta}(\lambda_k)$, but the converse is not true.
- Path algorithms for computing lasso solutions proceed by fitting models at a grid of λ values. The active set changes whenever λ crosses a knot, and predictor variables can both enter and leave the active set. However, at the first two knots λ_1 and λ_2 no variable can leave the active set. So the first two knots correspond to the first two variables entering the model.

Lockhart et al. prove that, under the null hypothesis that A_k contains all the strong predictor variables, the distribution of a test statistic $T_k \propto \lambda_k(\lambda_k - \lambda_{k+1})$ is asymptotically Exp(1). In the lasso case we know a lot about the knots and active set, but the group lasso picture is slightly more complicated. For the group lasso, $\hat{\beta}(\lambda)$ does not have piecewise linear components. To overcome this difficulty we will restrict our attention to the first group of variables to enter the active set since the analysis then follows almost exactly as for the lasso. (**To do:** Change this if I can find λ_k).

The *group lasso estimator* is the following solution to a convex problem

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{g=1}^{G} w_g \|\beta_g\|_2 \tag{2}$$

The parameter $\lambda \geq 0$ enforces sparsity in groups: for large λ most of the β_g will be zero vectors. The weights w_g are usually taken to be $\sqrt{p}_g$ to normalize the penalty across groups. Note that this includes the usual lasso estimator as a special case when all of the groups are of size 1, since then the penalty term is the L^1-norm of β.

To do: Fix references below

The group lasso estimator is discussed in (ref that guy's thesis) and (ref Yuan and Lin). An important extension is the sparse group lasso (ref ???) which enforces sparsity in groups as well as sparsity of the coefficients within the groups. For a survey on group lasso and related factor models see (ref???). **To do:** review some more literature and add a few more references here if they seem worthwhile.

Before considering the group lasso, we review some ingredients of the proof for the lasso. Let $J = \{1, 2, \ldots, p\}$ index variables and consider a stochastic process $f_{j,s} = sX_j^T y$ defined on $T = J \times [-1, 1]$. This stochastic process is simply a collection of linear combinations of y, hence it is Gaussian under the assumption of Gaussian errors. The *Karush-Kuhn-Tuckher (KKT) conditions* (ref???) imply that $\lambda_1 = \max_j |X_j^T y|$. By introducing the sign variable s, we can remove the absolute value and write λ_1 as the maximum of our Gaussian process

$$\lambda_1 = \max_{(j,s)} f_{j,s} \tag{3}$$

We have exhibited the first knot as the maximum of a Gaussian process. We can do this for the second knot by introducing a new process. Let (j_1, s_1) be the maximizer so that $\lambda_1 = s_1 X_{j_1}^T y$, and define

$$\begin{aligned} f_{(j,s)}^{(j_1,s_1)} &= \frac{sX_j^T y - sX_j^T X_{j_1} X_{j_1}^T y}{1 - ss_1 X_j^T X_{j_1}} \\ &= \frac{sX_j^T (I - P_{j_1}) y}{1 - s_1 X_{j_1}^T X_j s} \end{aligned} \tag{4}$$

where P_j is the projection onto the subspace spanned by X_j. We can think of this as a "residual process" after regressing out the maximum. Write $M = \max_{j \neq j_1, s} f_{(j,s)}^{(j_1,s_1)}$, the maximum of this residual process. It can be shown from the KKT conditions that $M = \lambda_2$. To summarize, we have represented the first two knots of the lasso solution path as the maxima of some natural Gaussian processes. Distributional facts about Gaussian processes now allow us to make conclusions about the distribution of functions of the knots.

3.2. Group lasso

To extend this argument to the group lasso we need to define Gaussian processes that characterize the knots of the group lasso solution path.

To do: Either use the simplified argument in the case of equal weights (equal group sizes), or finish adjusting the proof of $M = \lambda_2$ to include the weights and use that version. The proof can go in an appendix.

To do: Modify write-up to match notation with Taylor et al. (2013) and then paste it in here

3.3. Better p-values

Maybe just reference Taylor et al. (2013)? It might be worthwhile to leave in the connection with Lockhart et al. (2013) (getting our test statistic by not using the approximation)

- **To do:** Add the figures to this section
- **To do:** Convert to exposition instead of bulleted list
- **To do:** Update to reflect latest work
- As in LTTT, $T = \lambda_1(\lambda_1 - M)$ and $M = \lambda_2$
- Convergence to the limiting Exp(1) distribution is too slow

$$\frac{P(\chi_k / w_g \geq m + t/m)}{P(\chi_k / w_g \geq m)} \to e^{-t} \text{ as } m \to \infty$$

(when the group achieving λ_1 is group g and has rank k)

- The limiting distribution only depends on T, but we also observe M
- Let's just try the ratio (conditional χ_k tail probability) evaluated at T and M (it works better)
- Going one step further, instead of using the approximation (see LTTT Proof of Lemma 5)

$$\frac{M + \sqrt{M^2 + 4t}}{2} \approx M + \frac{t}{M}$$

  we can just use the left hand side
- For $T = \lambda_1(\lambda_1 - M)$ the left hand side simplifies to λ_1
- Now our p-value is

$$\frac{P(\chi_k/w_g \geq \lambda_1)}{P(\chi_k/w_g \geq \lambda_2)}$$

4. Simulations

To do: Expand, include simulation results (copy latex for includegraphics etc), do the simulation comparing stopping rules

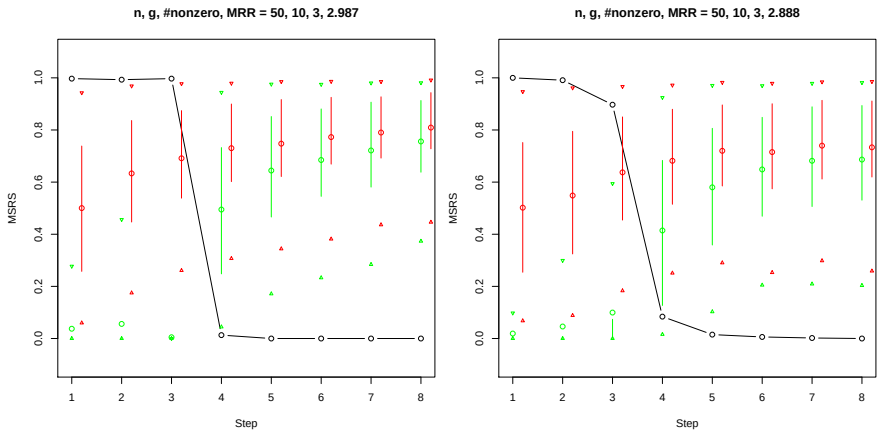
- Show null and non-null p-value-by-step plots for several examples
- Discuss one of these carefully (perhaps with a noise variable entering before a signal variable), for several fixed stopping points
- Compare a few stopping rules, e.g. naive one(s), TailStop/HybridStop, AIC/BIC/Cp
- (the stopping comparisons are not done yet, but I am guessing they will be comparable, perhaps AIC/BIC/Cp will be better, and in that case look for other advantages of ours to mention, e.g. interpretability for being based on p-values rather than a complexity penalty)

	Fdp	R	S	V	Power
(1) first	0.02	2.65	2.55	0.10	0.85
(1) forward	0.00	0.92	0.92	0.00	0.31
(1) hybrid	0.01	2.00	1.99	0.01	0.66
(2) first	0.03	2.54	2.41	0.13	0.80
(2) forward	0.00	0.95	0.95	0.00	0.32
(2) hybrid	0.00	2.00	1.99	0.01	0.66

TABLE 1

Evaluation of model selection using several stopping rules based on our p-values. The naive stopping rule performs well.

The large simulation has 50 groups, 25 of size 1, 10 of size 5, 10 of size 10, and 5 of size 15. Signal vectors were supported on random choices of 10 of these groups (changing in each realization), with non-zero signal magnitude around $\sqrt{2 \log p}$ where $p = 250$ (roughly 3.3). The design matrix had 200 rows and the simulation performed 100 realizations.



(a) Gaussian design (b) Categorical design

FIG 1. *The left panel shows results of a simulation with an independent Gaussian design matrix and a signal vector having three groups with non-zero coefficients of sizes one, two, and three, and seven groups of noise variables with sizes varying from one to ten, for a total of 30 variables. For the right panel we repeated this setup, but group sizes corresponded to levels of categorical variables. The groups of size one were increased to two, so there are 10 categorical variables with a total of 32 levels.*

5. Real data example

To do: Simulation with non-Gaussian design matrix (from AIDS data probably)

6. Discussion

To do: After finising everything else, move some points of discussion here, re-read and see if there's anything else interesting to comment on here. Mention ongoing work

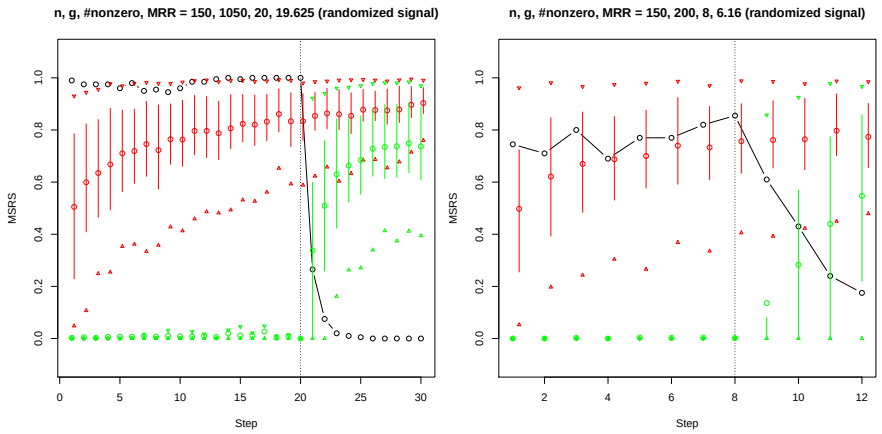
Appendix A: Derivation of closed forms of statistic

For the sake of completeness this appendix provides full derivations of closed forms for the quantities required to compute our statistic. We require two facts from Taylor et al. (2013)

A point $\eta \in \mathcal{K}$ maximizes f_η over a convex set $\mathcal{K}$ if and only if the following conditions hold:

$$\nabla f_{|T_\eta \mathcal{K}} = 0, \quad \widetilde{f}_\eta^\eta \geq \mathcal{V}_\eta^+, \quad \widetilde{f}_\eta^\eta \leq \mathcal{V}_\eta^-, \quad \text{and} \quad \mathcal{V}_\eta^0 \leq 0. \tag{5}$$

The same equivalence holds true even when $\mathcal{K}$ is only locally convex.



(a) Large Gaussian lasso (b) Large Gaussian group lasso

FIG 2. *The left panel shows results of a simulation with an independent Gaussian design matrix and a signal vector having three groups with non-zero coefficients of sizes one, two, and three, and seven groups of noise variables with sizes varying from one to ten, for a total of 30 variables. For the right panel we repeated this setup, but group sizes corresponded to levels of categorical variables. The groups of size one were increased to two, so there are 10 categorical variables with a total of 32 levels.*

Write $M^{\pm}_{\eta,h}$ and $M^{0}_{\eta,h}$ as the corresponding suprema and infima over the restricted parameter set S_h. Note that the characterization of the global maximizer (Lemma 1, general paper) implies $M^{0}_{\eta,h} \leq 0$ and $M^{+}_{\eta,h} \leq M^{-}_{\eta,h}$ for all $h \neq g$. This will be used below to eliminate some degenerate cases of the optimization sub-problem on each S_h.

Now fix h, let $P_g = X_g X_g^T$ and define

$$\begin{aligned} a_h &= w_h^{-1} X_h^T (I - P_g) y \\ b_h &= w_g w_h^{-1} X_h^T P_g y / \|X_g^T y\| \\ c_h &= a_h^T b_h / (\|a_h\| \|b_h\|) \\ K_h &= \{x : \|x\| = 1, b_h^T x > w\} \end{aligned}$$

Note that $K_h = \emptyset$ if and only if $\|b_h\| < w$. We will consider two cases. First, suppose $\|b_h\| > w$. In this case we must rule out one sub-case, when a_h is not in the polar cone of K_h. If this occurs, then on at least one side of the boundary of K_h we have $a_h^T x > 0$ there, so that the infimum inside K is $-\infty$ but the supremum outside K is $+\infty$. This violates the characterization of the global maximizer of the original process as described above, so this case cannot occur.

Now, if a_h is in the polar cone of K_h (and ruling out the probability zero case where $a_h^T x = 0$ on the boundary of K_h), then the numerator $a_h^T x < 0$ on K_h so there is a finite positive infimum in the interior K_h. Similarly there is a

	Fdp	R	S	V	Power
(2) first	0.20	9.64	7.47	2.17	0.93
(2) forward	0.26	1.00	0.74	0.26	0.09
(2) hybrid	0.27	2.00	1.46	0.55	0.18
(3) first	0.32	9.32	6.46	2.86	0.81
(3) forward	0.64	0.98	0.34	0.64	0.04
(3) hybrid	0.50	2.00	0.99	1.00	0.12

TABLE 2

Evaluation of model selection using several stopping rules based on our p-values. The naive stopping rule performs well.

finite positive supremum on the interior of K_h^c (the numerator is negative near the boundary of K_h by continuity).

To attain the infimum on K_h requires making x close to b_h and simultaneously making the numerator closer to zero, so x should be on the side of b_h that is closer to a_h. To attain the supremum on K_h^c requires making x simultaneously close to a_h (so that $a_h^T x > 0$) and b_h (to make the denominator small). In summary, both the infimum and supremum are attained with x between a_h and b_h, when the angle θ between a_h and b_h is larger than both the angle ψ between a_h and x and the angle ϕ between x and b_h. This allows the simplification $\phi = \theta - \psi$. This is easier to verify for the case when $\|b_h\| < w$.

We have verified that solving both cases requires finding the maxima of the trigonometric form $\|a_h\| \cos(\psi)/(w - \|b_h\| \cos(\theta - \psi))$ of the linear fraction on the interiors of K_h (if it is nonempty) and K_h^c. The derivative is proportional to

$$\frac{\|b_h\| \sin(\theta) - w \sin(\psi)}{w - \|b_h\| \cos(\theta - \psi)}$$

Since we have ruled out the boundary of K_h, we can ignore the denominator and focus on critical points corresponding to the angles $\psi^{\pm}$ (symmetric about $\pi/2$) where $\sin(\psi^{\pm}) = (\|b_h\|/w) \sin(\theta)$. That is,

$$\psi^{\pm} \in \arcsin \frac{\|b_h\|}{w} \sqrt{1 - c_h^2}$$

Note that these critical points only exist if $\sin(\theta) < w/\|b_h\|$, and since we have argued that the infimum and supremum are attained it follows that this condition is met in our case. Let ψ^+ denote the smaller of the two angles. If K_h is nonempty, then ψ^- gives the infimum in K_h and ψ^+ the supremum on K_h^c. This is necessary since if x is in K_h then its angle from a_h is larger than $\pi/2$ because a_h is in the polar cone of K_h. If K_h is empty, then the supremum is still attained at ψ^+ since the numerator is negative at $\psi^- > \pi/2$ and the denominator is always positive.

Finally we calculate the values of the linear fraction at $\psi^{\pm}$. Let us first record

some facts:

$$0 < \psi^+ < \pi/2, \quad \psi^- = \pi - \psi^+$$
$$\cos(\theta) = c_h, \quad \sin(\theta) = \sqrt{1 - c_h^2}$$
$$\sin(\psi^+) = \sin(\psi^-) = \frac{\|b_h\|}{w} \sin(\theta)$$
$$\cos(\psi^+) = -\cos(\psi^-) = \sqrt{1 - (\|b_h\|/w)^2(1 - c_h^2)}$$

Using the angle difference formula,

$$\begin{aligned}\cos(\theta - \psi^+) &= \frac{\|b_h\|}{w}(1 - c_h^2) + c_h \cos(\psi^+) \\ &= \frac{\|b_h\|}{w}(1 - c_h^2) + c_h \cos(\psi^+)\end{aligned}$$

Hence

$$M^+_{\eta,h} = \frac{\|a_h\| c_h}{w - (\|b_h\|^2/w)(1 - c_h^2) - \|b_h\| c_h \cos(\psi^+)}$$

And $M^-_{\eta,h}$ is ∞ when $\|b_h\| < w$ and otherwise given by

$$M^+_{\eta,h} = \frac{\|a_h\| c_h}{w - (\|b_h\|^2/w)(1 - c_h^2) + \|b_h\| c_h \cos(\psi^+)}$$

I haven't been able to algebraically simplify these yet to a form that allows my λ_2 proof below to go through. The version below just assumes $\|b_h\| < w = 1$.

We can rewrite this by rationalizing the denominator:

$$M = \frac{a_h^T b_h + \sqrt{a_h^T a_h (1 - b_h^T b_h) + (a_h^T b_h)^2}}{1 - b_h^T b_h}$$

Leaving M in this form we now consider λ_2. The KKT conditions for the group lasso problem give (directly from Yuan and Lin)

$$\|X_g^T(y - X\beta)\| \le \lambda w_g \quad \forall \beta_g \neq 0$$
$$\beta_g = \left(1 - \frac{\lambda w_g}{\|S_g\|}\right)_+ S_g$$

where $S_g = X_g^T(y - X\beta_{-g})$. These expressions simplify for $\lambda_2 \le \lambda < \lambda_1$. Letting g be the index of the first group to enter we now have $S_g = X_g^T y$

$$\beta_g = \left(1 - \frac{\lambda w_g}{\|X_g^T y\|}\right) X_g^T y$$
$$\lambda w_h \ge \|X_h^T y - X_h^T P_g y (1 - \lambda w_g / \|X_g^T y\|)\| \quad \forall h \neq g$$

Now let $\lambda = \lambda_2$, so the inequality above is strict for all other groups and equality is attained for the second group to enter. Hence λ_2 satisfies

$$\begin{aligned}\lambda_2 &= \max_{h \neq g} \|X_h^T(I - P_g)y - \lambda_2(w_g X_h^T P_g y / \|X_g^T y\|)\|/w_h \\ &= \max_{h \neq g} \|a_h - \lambda_2 b_h\|\end{aligned}$$

Write $\lambda_{2,h}$ as the positive root of the quadratic equation obtained by squaring the norm, so $\lambda_{2,h}^2 = \|a_h - \lambda_2 b_h\|^2$. Then $\lambda_2 = \max_{h \neq g} \lambda_{2,h}$. Solving this (note that $b_h^T b_h < 1$), we find

$$\lambda_{2,h} = \frac{a_h^T b_h + \sqrt{a_h^T a_h (1 - b_h^T b_h) + (a_h^T b_h)^2}}{1 - b_h^T b_h}$$

Hence when h is the index of the second group to enter, we have $\lambda_2 = \lambda_{2,h} = M$.

References

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2013). A significance test for the lasso. *arXiv:1301.7161*. Submitted to Annals of Statistics, URL `http://arxiv.org/abs/1301.7161`.

Ming, Y. and Lin, Y. (2005). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, **68** 49–67.

Taylor, J., Loftus, J. and Tibshirani, R. (2013). Tests in adaptive regression via the kac-rice formula. *ArXiv e-prints*. `1308.3020`.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, **58** 267–288.